Inferring Gene Regulatory Networks from Time-Ordered Gene Expression Data Using Differential Equations

Michiel de Hoon, Seiya Imoto, and Satoru Miyano

Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
{mdehoon,imoto,miyano}@ims.u-tokyo.ac.jp

Abstract. Recently, cDNA microarray experiments have generated large amounts of gene expression data. In time-ordered gene expression data, the expression levels are measured at several points in time following some experimental manipulation. A gene regulatory network can be inferred by describing the gene expression data in terms of a linear system of differential equations. As biologically the gene regulatory network is known to be sparse, we expect most coefficients in such a linear system of differential equations to be zero. In previously proposed methods to infer a linear system of differential equations, some ad hoc assumptions are made to limit the number of nonzero coefficients in the system. Instead, we propose to infer the degree of sparseness of the gene regulatory network from the data, where we determine which coefficients are nonzero by using Akaike's Information Criterion.

1 Introduction

The recently developed cDNA microarray technology allows gene expression levels to be measured for the whole genome at the same time. While the amount of available gene expression data has been increasing rapidly, the required mathematical techniques to analyse such data is still in development. Particularly, deriving a gene regulatory network from gene expression data has proven to be a difficult task.

In time-ordered gene expression measurements, the temporal pattern of gene expression is investigated by measuring the gene expression levels at a small number of points in time. Periodically varying gene expression levels have for instance been measured during the cell cycle of the yeast *Saccharomyces Cerevisiae* [1]. The gene response to a slowly changing environment has been measured during the diauxic shift of the same yeast [2]. In other experiments, the temporal gene expression pattern due to an abrupt change in the environment of the organism is measured. As an example, the gene expression response was measured of the cyanobacterium *Synechocystis* sp. PCC 6803 after to sudden shift in the intensity of external light [3].

A number of methods have been proposed to infer gene interactions from gene expression data. In cluster analysis [2, 4, 5], genes are grouped together

based on the similarity between their gene expression profiles. Several measures of sensitivity can be used, such as the Euclidean distance, correlation, or angle between two vectors containing the gene expression data. Inferring Boolean or Bayesian networks from measured gene expression data has been proposed previously [6–9], as well as modelling gene expression data using an arbitrary system of differential equations [10]. However, a long series of time-ordered gene expression data would be needed to reliably infer such an arbitrary system of differentional equations. This is currently often not yet available.

Instead, we will consider inferring a linear system of differential equations from gene expression data. This approach maintains the advantages of quantitativeness and causality inherent in differential equations, while being simple enough to be computationally tractable.

Previously, modelling biological data with linear differential equations was considered theoretically by Chen [11]. In this model, both the mRNA and the protein concentrations were described by a system of linear differential equations. Such a system can be described as

$$\frac{\mathrm{d}}{\mathrm{d}t}\underline{x}\,(t) = \underline{\underline{M}} \cdot \underline{x}\,(t)\,, \tag{1}$$

in which $\underline{\underline{M}}$ is a constant matrix with units of $[\mathrm{second}]^{-1}$, and the vector $\underline{x}\,(t)$ contains the mRNA and protein concentrations as a function of time. To infer the coefficients in the system of differential equations from measured data, two methods were suggested.

The first method expands the measured gene expression data in a Fourier series with a limited number of Fourier components. This number was chosen to be equal to the number of time points at which the gene expression level was measured. Using more Fourier components would lead to an underdetermined system.

The second method is to discretise the system of differential equations, substitute the measured mRNA and protein concentrations, and to solve the resulting linear system of equations in order to find the coefficients in the system of linear differential equations. The system is simplified by assuming that the concentration of one protein does not affect other proteins directly, and similarly for genes, in addition to the assumption that one type of mRNA is involved in the production of one type of protein only. However, the resulting system of equations is still underdetermined. Using the additional requirement that the gene regulatory network should be sparse, it is shown that the model can be constructed in $O\left(m^{h+1}\right)$ time, where m is the number of genes and h is the number of non-zero coefficients allowed for each differential equation in the system [11]. The parameter h is chosen ad hoc.

Although describing a gene regulatory network with differential equations is appealing, there is one drawback of the method proposed by Chen. For a given parameter h, each column in the matrix $\underline{\underline{M}}$ will have exactly h nonzero elements. This means that every gene or protein in the system affects h other genes or proteins. This has two consequences:

- No genes or proteins can exist at the bottom of a network. Every gene and protein is the parent of h other genes or proteins in the network.
- The network that is found inevitably contains loops.

While feedback loops are likely to exist in gene regulatory networks, the loops that are found using this method are artificially produced by the method itself. The method dictates the existance of loops. Instead, we would like to determine the existance of loops in the gene regulatory network from the data.

At the other extreme, no loops are allowed in Bayesian networks. Bayesian networks rely on the joint probability distribution of the estimated network being able to be decomposed in a product of conditional probability distributions. This decomposition is possibly only in the absence of loops. In addition, Bayesian networks tend to contain many parameters, and therefore a large amount of data is needed to estimate such a model.

We therefore aim to find a method that can allows the existance of loops in the network, but does not dictate their presence. Using equation (1), we also construct a sparse matrix by limiting the number of non-zero coefficients that may appear in the system. However, we do not choose this number ad hoc; instead, we estimate the number of nonzero parameters from the data by using Akaike's Information Criterion (AIC). This enables us to obtain the sparsity of the gene regulatory network from the gene expression data. In contrast to previous methods, the number of gene regulatory pathways is allowed to be different for each gene.

Usually, in cDNA microarray experiments only the gene expression levels are found by measuring the corresponding mRNA concentrations, whereas the protein concentration is unknown. To analyze the results from such experiments, we therefore construct a system of differential equations in which genes are allowed to affect each other directly, since proteins are no longer available in the model to act as an intermediary. The vector $\underline{x}$ then contains only the mRNA concentrations only, and matrix $\underline{\underline{M}}$ describes gene-gene interactions only.

2 Method

Consider the gene expression ratios of m genes as a function of time. At a given time t, the expression ratios can be expressed as a vector $\underline{x}\,(t)$ with m entries. The interactions between these genes can be described quantitatively in terms of a system of differential equations. Several forms can be chosen for the differential equations. We have chosen a system of linear differential equations (1), which is the simplest possible model. This equation can be solved as

$$\underline{x}\,(t) = \exp\left(\underline{\underline{M}}t\right) \cdot \underline{x}_0, \tag{2}$$

in which $\underline{x}_0$ is the gene expression ratio at time zero. In this equation, the matrix exponential is defined by the Taylor expansion of the exponential function [12]:

$$\exp\left(\underline{\underline{A}}\right) \equiv \sum_{i=0}^{\infty} \frac{1}{i!} \underline{\underline{A}}^i. \tag{3}$$

Since equation (2) is nonlinear in $\underline{\underline{M}}$, it will still be very difficult to solve for $\underline{\underline{M}}$ using experimental data. We therefore approximate the differential equation (1) by its discretized form:

$$\frac{\Delta \underline{x}}{\Delta t} = \underline{\underline{M}} \cdot \underline{x} \; . \tag{4}$$

We can now substitute a given set of time-ordered gene expression data $\underline{x}_{t_i}$ at times t_i, $i \in \{1, \ldots, n\}$. We find

$$\underline{x}_{t_i} - \underline{x}_{t_{i-1}} = (t_i - t_{i-1}) \cdot \underline{\underline{M}} \cdot \underline{x}_{t_{i-1}} \; . \tag{5}$$

To this equation we can add a error ε_{t_i}, which will invariably be present in the data:

$$\underline{x}_{t_i} - \underline{x}_{t_{i-1}} = (t_i - t_{i-1}) \cdot \underline{\underline{M}} \cdot \underline{x}_{t_{i-1}} + \underline{\varepsilon}_{t_i} \; . \tag{6}$$

By using this equation, we effectively describe a gene expression network in terms of a multidimensional linear Markov model.

We assume that the measurement error has a normal distribution independent of time:

$$f\left(\underline{\varepsilon}_t; \sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \exp\left\{-\frac{\underline{\varepsilon}_t^{\mathrm{T}} \cdot \underline{\varepsilon}_t}{2\sigma^2}\right\} \; , \tag{7}$$

with a standard deviation σ equal for all genes at all times. The log-likelihood function for a series of measurements at n time points is then

$$L\left(\underline{\underline{M}}, \sigma^2\right) = -\frac{nm}{2} \ln\left[2\pi\sigma^2\right] - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \hat{\underline{\varepsilon}}_{t_i}^{\mathrm{T}} \cdot \hat{\underline{\varepsilon}}_{t_i} \; , \tag{8}$$

in which

$$\hat{\underline{\varepsilon}}_{t_i} = \underline{x}_{t_i} - \underline{x}_{t_{i-1}} - (t_i - t_{i-1}) \cdot \underline{\underline{M}} \cdot \underline{x}_{t_{i-1}} \tag{9}$$

is the measurement error at time t_i estimated from the measured data.

The maximum likelihood estimate of the variance σ^2 can be found by maximising the log-likelihood function with respect to σ^2. This yields

$$\hat{\sigma}^2 = \frac{1}{nm} \sum_{i=1}^{n} \hat{\underline{\varepsilon}}\left(t_i\right)^{\mathrm{T}} \cdot \hat{\underline{\varepsilon}}\left(t_i\right) . \tag{10}$$

Substituting this into the log-likelihood function (8) yields

$$L\left(\underline{\underline{M}}, \sigma^2 = \hat{\sigma}^2\right) = -\frac{nm}{2} \ln\left[2\pi\hat{\sigma}^2\right] - \frac{nm}{2}. \tag{11}$$

The maximum likelihood estimate $\hat{\underline{\underline{M}}}$ of the matrix $\underline{\underline{M}}$ can now be found by minimizing $\hat{\sigma}^2$. By taking the derivative of equation (10) with respect to $\underline{\underline{M}}$, we find that $\hat{\sigma}^2$ is minimized for

$$\hat{\underline{\underline{M}}} = \underline{\underline{A}}^{-1} \cdot \underline{\underline{B}} \; , \tag{12}$$

where the matrices $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are defined as

$$\underline{\underline{A}} \equiv \sum_{i=1}^{n} \left[(t_i - t_{i-1})^2 \cdot \underline{x}_{i-1} \cdot \underline{x}_{i-1}^{\mathrm{T}} \right] \tag{13}$$

and

$$\underline{\underline{B}} = \sum_{i=1}^{n} \left[(t_i - t_{i-1}) \cdot \left(\underline{x}_i - \underline{x}_{i-1} \right) \cdot \underline{x}_{i-1}^{\mathrm{T}} \right]. \tag{14}$$

In the absence of errors, the estimated matrix $\underline{\underline{\hat{M}}}$ is equal to the true matrix $\underline{\underline{M}}$. We know from biology that the gene regulatory network and therefore $\underline{\underline{M}}$ is sparse. However, the presence of noise in experiments would cause most or all of the elements in the estimated matrix $\underline{\underline{\hat{M}}}$ to be nonzero, even if the corresponding element in the true matrix $\underline{\underline{M}}$ is zero. We can determine if a nonzero matrix element is due to noise by setting it equal to zero and recalculating the total squared error as given in equation (10). If the increase in the total squared error is small, we conclude that the previously calculated value of the matrix element is due to noise.

Formally, we can decide if matrix elements should be set to zero using Akaike's Information Criterion [13, 14]

$$AIC = -2 \cdot \begin{bmatrix} \text{log-likelihood of the} \\ \text{estimated model} \end{bmatrix} + 2 \cdot \begin{bmatrix} \text{number of estimated} \\ \text{parameters} \end{bmatrix} \tag{15}$$

in which the estimated parameters are $\hat{\sigma}^2$ and the elements of the matrix $\underline{\underline{\hat{M}}}$ that we allow to be nonzero. The AIC avoids overfitting of a model to data by comparing the total error in the estimated model to the number of parameters that was used in the model. The model which has the lowest AIC is then considered to be optimal. The AIC is based on information theory and is widely used for statistical model identification, especially for time series model fitting [15].

Substituting the estimated log-likelihood function from equation (11) into equation (16), we find

$$AIC = nm \ln \left[2\pi\hat{\sigma}^2 \right] + nm + 2 \cdot \begin{bmatrix} \text{number of nonzero} \\ \text{elements in } \underline{\underline{\hat{M}}} \end{bmatrix}. \tag{16}$$

From this equation, we see that while the squared error increases, the AIC may decrease as the number of nonzero elements decreases. A gene regulatory network can now be estimated using the following procedure. Starting from the measured gene expression levels $\underline{x}_i$ at time points t_i, we calculate the matrices A and B as defined in equations (13) and (14). We find the maximum likelihood estimate $\underline{\underline{\hat{M}}}$ of the matrix $\underline{\underline{M}}$ from equation (12). The corresponding squared error is found from equations (9) and (10). Equation (16) gives us the AIC for the maximum likelihood estimate of $\underline{\underline{M}}$. We then generate a new matrix $\underline{\underline{\hat{M}}}'$ from $\underline{\underline{\hat{M}}}$ by setting a set of matrix elements equal to zero. We recalculate the squared error and the

AIC for this modified matrix $\underline{\underline{\hat{M}}}'$. The matrix $\underline{\underline{\hat{M}}}'$, and a corresponding set of zeroed matrix elements, with the lowest AIC is then the final estimated gene regulatory network.

In typical cDNA microarray experiment, the number of genes is several thousands. Usually only a small number of genes (several tens to hundreds) are affected by the experimental manipulation. The matrix $\underline{\underline{M}}$ is therefore quite large, and the number of sets of zeroed matrix elements is extremely large. An exhaustive search to find the optimal combination of zeroed matrix elements is therefore not feasible. Instead, we propose a greedy search. First, we randomly choose an initial set of matrix elements that we set equal to zero. For every matrix element, we determine if the AIC is reduced if we change the state of the matrix element between zeroed and not zeroed. If the AIC is reduced, we change the state of the matrix element and continue with the next matrix element. This process is stopped if the AIC can no further be reduced. We repeat this algorithm many times starting from different initial sets of zeroed matrix elements. If the algorithm described above yields the same set of zeroed elements several times, we can assume that no other sets of zeroed elements with a lower AIC exist.

3 Discussion

We have shown a method to derive a gene regulatory network in the form of a linear system of differential equations from measured gene expression data. Due to the limited number of time points at which measurements are typically made, finding a gene regulatory network is usually an underdetermined problem. Since in biology the resulting gene regulatory network is expected to be sparse, we set most of the matrix entries equal to zero, and derived a network using only the nonzero entries. The number of nonzero entries, and therefore the sparsity of the network, was derived from the data using Akaike's Information Criterion.

Describing a gene network in terms of differential equations has three advantages. First, the set of differential equations describes causal relations between genes: a coefficient M_{ij} of the coefficient matrix determines the effect of gene j on gene i. Second, it describes gene interactions in an explicitly numerical form. Third, because of the large amount of information present in a system of differential equations, other network forms can easily be derived from it. We can also link the inferred network to other analysis or visualisation tools, for instance *Genomic Object Net* [16].

In previously described methods to derive gene regulatory networks from gene expression data, either loops cannot be found (Bayesian networks) or the method artificially generates loops in the network. While the method proposed here allows loops to be present in the network, it does not require their existence. Loops are only found if the measured data warrant them. We aim to apply our method to measured gene expression data to evaluate its usefulness in practice.

References

1. Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell **9** (1998) 3273–3297.
2. DeRisi, J., Iyer, V., Brown, P.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278** (1997) 680–686.
3. Hihara, Y., Kamei, A., Kanehisa, M., Kaplan, A., Ikeuchi, M.: DNA microarray analysis of cyanobacterial gene expression during acclimation to high light. The Plant Cell **13** (2001) 793–806.
4. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95** (1998) 14863–14868.
5. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T.: Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA **96** (1999) 2907–2912.
6. Liang, S., Fuhrman, S., Somogyi, R.: REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Proc. Pac. Symp. on Biocomputing **3** (1998) 18–29.
7. Akutsu, T., Miyano, S., Kuhara, S.: Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics **16** (2000) 727–734.
8. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. J. Comp. Biol. **7** (2000) 601–620.
9. Imoto, S., Goto, T., Miyano, S.: Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. Proc. Pac. Symp. on Biocomputing **7** (2002) 175–186.
10. Sakamoto, E., Iba, H.: Evolutionary inference of a biological network as differential equations by genetic programming. *Genome Informatics* **12** (2001) 276–277.
11. Chen, T., He, H., Church, G.: Modeling gene expression with differential equations. Proc. Pac. Symp. on Biocomputing **4** (1999) 29–40.
12. Horn, R., Johnson, C.: Matrix Analysis. Cambridge University Press, Cambridge, UK (1999).
13. Akaike, H.: Information theory and an extension of the maximum likelihood principle. Research Memorandum No. 46, Institute of Statistical Mathematics, Tokyo (1971). In Petrov, B. and Csaki, F. (editors): 2nd Int. Symp. on Inf. Theory. Akadémiai Kiadó, Budapest (1973) 267–281.
14. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automat. Contr. **AC-19** (1974) 716–723.
15. Priestley, M.: Spectral Analysis and Time Series. Academic Press, London (1994).
16. Matsuno, H., Doi, A., Hirata, Y., Miyano, S.: XML documentation of biopathways and their simulation in Genomic Object Net. *Genome Informatics* **12** (2001) 54–62.